Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 31 (2014) 663 – 670
Procedia Computer Science

2nd International Conference on Information Technology and Quantitative Management, ITQM 2014

Feature Extension for Short Text Categorization Using Frequent Term Sets

Yuan Man[a,b]*

[a]*China Huarong Asset Management CO., LTD., Beijing, China*
[b]*Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences , Univiersity of Chinese Academy of Sciences, Beijing, China*

Abstract

A short text feature extension method based on frequent term sets is proposed to overcome the drawbacks of the vector space model (VSM) on representing short text content. After defining the co-occurring and class orientation relations between terms, frequent term sets with identical class orientation are generated by calculating the support and confidence of word sets, and then taken as the background knowledge for short text feature extension. For each single term of the short text, the term sets containing this term are retrieved in the background knowledge and added into the original term vector as the feature extension. The experimental results on Sougou corpus show that the support and confidence have great impact on the scale of the background knowledge, but excessive extension also has redundancy and cannot obtain further improvement. The background knowledge based on frequent term sets is an effective way for feature extension. When the number of the training documents is limited, these extended features can greatly improve the classification results of SVM.

Keywords: frequent term sets, short text classification, feature extension

1. Main text

Text representation which refers to converting text content to a certain format for computer to process is a fundamental problem in text mining. The most popular text representation method in present is Vector Space

* Corresponding author. Tel.: +86-010-59618543; fax: +86-010-59618543.
E-mail address: yuanman@chamc.com.cn.

doi:10.1016/j.procs.2014.05.314

Model (VSM) [1]. In VSM, text contents are treated as Bag of Words (BOW), which ignores the associations of text features as well as the context and grammar structure. In recent years, following the emerging internet media such as social network and micro-blog, short text becomes an important type of text content online. Since the terms are rare in short text, it is weak for describing specific information, which aggravates the limits of VSM. To solve this problem, one of the method is to extend the text features to include more semantic, context and associations, using text mining technologies such as natural language processing, background knowledge and frequent item set mining.

Feature extension based on natural language processing is to use language models, grammar and syntax analysis to construct complex feature units [2-4]. Feature extension based on background knowledge tries to obtain more semantic information by retrieve the term in search engines [5,6], Wikipedia [7,8] or other outer resources. Frequent item set is a concept in association rule mining, and it also means frequent term set in text mining. It has been applied in text mining [9] because frequent term set reflect the associations of terms which involve more context and semantic information than single words. Cheng [10] analyzed frequent term sets for text categorization and proposed a frequent term sets selection method using multi-restraint metrics such as relevancy, coverage and redundancy. Ahone [11] firstly discussed the maximum frequent sequence (MFS) mining algorithm to avoid the duplication problem between each term set. Hernández [12] use MFS for text representation and each MFS corresponds to a text feature in the text vector space. A text clustering algorithm is then applied on this feature space.

This paper aims to improve short text representation in vector space model for short text categorization. A feature extension method based on background knowledge using frequent term sets is presented. Firstly, we define the co-occurring and class orientation relation of terms. Double term sets with co-occurring relation and identical class orientation relation are extracted from long text content to build the background knowledge. Then the original features are extended by the background knowledge and new features are added into the new feature space on which the SVM classifier is trained. Finally, experimental evaluation on real text data is conducted.

2. Feature extension based on background knowledge

Short text on Internet include multiple forms such as search key word, web review, micro-blog and news title. In this paper, we mainly focus on news title which appears frequently on social network feeds and shares, and the background knowledge are extracted from news content. The procedures involve: (1) text pre-processing, such as stemming and word segmentation; (2) mining double term sets to build the background knowledge; (3) feature extension using background knowledge and SVM classifier training; (4) feature extension on test data and evaluation of classification result.

2.1. Background knowledge based on frequent term sets

Background knowledge are extracted from full content of document set $D=\{d_1, d_2, \ldots, d_n\}$ which is relevant to the short text. In document set D, term set $T=\{t_1, t_2, .., t_k\}$ is the collection of k terms, and $C=\{c_1, c_2, \ldots, c_m\}$ is the collection of class labels.

To select valuable frequent term sets, we consider the following restraints:

Definition 1 (Support): The support of term set T is the number of documents which contain T dividing the number of all the documents in data set , formulated as $sup(T) = Count(D_T) / Count(D)$. D_T is the number of documents which contain T and $Count(D_T)$ is the number of documents in D_T and $Count(D)$ is the number of documents in D.

Definition 2 (Confidence): The confidence of association rule $t \Rightarrow c$ is *conf(t,c)*, which means the number of documents containing t in class c dividing the number of all documents involving t, formulated as

$conf(t,c) = Count(D_{(t,c)}) / Count(D_t)$. D_T is the number of documents which contain T, and $D_{t,c}$ is the collection of documents in class c containing term t . *conf(t,c)* reflects the connection of term t and class c.

Definition 3 (Co-occurring relation): If the support of term set T surpasses a threshold α ($0 < \alpha \le 1$), T is a frequent term set and all terms in T have a Co-occurring relation.

Definition 4 (Class orientation): For term t and class $c \in C$, if *conf(t,c)* surpasses a threshold β ($0.5 \le \beta \le 1$), term t has a Class orientation to c, formulated as Tendency(t)=c.

Definition 5 (Identical Class Orientation): For two terms $t_1$and t_2, if there is a class c , Tendency(t_1)=c and Tendency(t_2)=c, $t_1$and t_2 have an Identical Class Orientation.

The extraction of background knowledge from full text content must fulfil the following two conditions: (1) The terms in a frequent term set have a Co-occurring relation which means the support of frequent term set is larger than the threshold α ; (2) The terms in a frequent term set also have an Identical Class Orientation which means there is a class c and all terms in the frequent term set has a Class orientation to c.

Terms with Identical Class Orientation can be deemed as terms from the same or close topic, and it is expected to possess better discriminative ability when extending features from these terms. Considering that most of phases or combinations of words in Chinese consist of 2 words, although there are term sets with more items, under the restraint of above metrics, the numbers are too rare to make substantial influences. So, the background knowledge that we extract are double term set from full text content. The detail of this algorithm can be described as:

Algorithm 1: Double term set extraction for background knowledge

Input: Feature set F, support threshold α , confidence threshold β , and class labels set C

Output: Double term set S

1. Initialize $S \leftarrow \theta$
2. for(i=1; i<=|F|; i++)
3. if(sup(t_i)>= α and c=Tendency(t_i) $\in C_k$)
4. $C_k \leftarrow C_k \cup \{t_i\}$
5. end if
6. end for
7. for(k=1;k<=|C|;k++)
8. for(i=1;i<=| C_i |;i++)
9. for(j=i+1;j<=| C_i |;j++)
10. if(sup(t_i, t_j) >= α) $S \leftarrow S \cup \{(t_i, t_j)\}$
11. end for
12. end for
13. end for
14. return S

The double term set extraction adopted Apriori algorithm. Since the term sets without Identical Class Orientation are not added into the background knowledge, the above algorithm can be improved by condition (2): firstly, compute all the class orientation and support, if there is no class orientation or inadequate support, discard the term; otherwise, divide these terms into different classes by their class orientations. Secondly, extract double term sets in each class.

2.2. Feature extension using background knowledge

The procedure of feature extension using background knowledge is: for each original term in short text, search for the term set in background knowledge which contains this term, and then use the pair word as extended feature to add into the original text vector. For example, assume feature vector of a text document is $\boldsymbol{d}$=$\{t_1, t_2, ..., t_k\}$, after searching for the pair word of each term and get the extended features $\{t_i\ t_{i+1}, ..., t_m\}$, the

new feature vector of document d will be $\{t_1, t_2, ..., t_k, t_i t_{i+1} ... t_m\}$. The details of this algorithm can be described as:

Algorithm 2 Short text feature extension

Input: Short text document $\boldsymbol{d}=\{t_1, t_2, ..., t_k\}$, background knowledge S

Output: Extended short text ***exd_d***

1. Initialize $\boldsymbol{exd\_d} \leftarrow \theta$
2. for(int m=1;i<=k;i++)
3. for each $(t_i, t_j) \in S$ do
4. if $t_m = t_i$ and t_j not in $\boldsymbol{d} \cup \boldsymbol{exd\_d}$, do
5. $\boldsymbol{exd\_d} \leftarrow \boldsymbol{exd\_d} \cup \{t_j\}$
6. end if
7. end for each
8. end for
9. $\boldsymbol{exd\_d} \leftarrow \boldsymbol{d} \cup \boldsymbol{exd\_d}$
10. return ***exd_d***

3. Experiment

3.1. Dataset

Sougou Dataset [13] is a collection of news articles provided by Sougou Lab, including 18 categories such as International, Sports, Society, Entertainment etc.. Each article document involve page URL, page ID, news title and page con-tent. The html tags have been removed in the page content. Our experiment selected 9 categories from the dataset and each category contains 4000 documents. Short text refers to the news title and the news contents are used for background knowledge extraction. All documents were pre-processed by word segmentation using ICTCLAS [14].

3.2. Experiment Results

The experiment first evaluated the influence of support and confidence threshold on the doublet term extraction and classification result; then, the classification results of original features and extended features were compared; finally, we discussed the effectiveness of feature extension on different scale f training documents. The feature selection method used here is Information Gain, feature weight is term frequency, and the classifier is Support Vector Machine (SVM).

The extraction of double term sets highly relies on support and confidence restraints. The support metric guarantees the co-occurring relation of terms and the confidence metric determines if the terms have identical class orientation. We extracted double term sets using α =1%, 1.5%, 2%，2.5%, and β =0.5, 0.6, 0.7, 0.8, 0.9, and the results are listed in Table 1.

Table 1. Number of extracted frequent term sets

α \ β	0.5	0.6	0.7	0.8
1%	11842	8005	4156	1758
1.5%	3857	2335	1196	398
2%	1929	1098	436	153
2.5%	985	655	296	80

As is shown in Table 1, when support or confidence arose, the number of extracted double term sets decreased rapidly. For example, then number of extracted double term sets when $\alpha = 1\%$, $\beta = 0.5$ is about 12 times of that when $\alpha = 2.5\%$, $\beta = 0.5$,and about 470 times of that when $\alpha = 2.5\%$, $\beta = 0.9$. It can be deduced here that when support and confidence are set to be high, the number of extracted term sets to be background knowledge will be too rare to have substantial influence on the original feature space. So, in the following experiment, we select 4 metric settings t build background knowledge: ($\alpha = 1\%$, $\beta = 0.5$), ($\alpha = 1\%$, $\beta = 0.6$), ($\alpha = 1\%$, $\beta = 0.7$) and ($\alpha = 1.5\%$, $\beta = 0.5$).

Table 2 is the classification result of the 4 metric settings. The classifier is SVM, and the ratio of training documents to test documents is 3:7 and 7:3. As we can observe, ($\alpha = 1\%$, $\beta = 0.5$) and ($\alpha = 1\%$, $\beta = 0.6$) can obtain better results than the other two settings. Additionally, although ($\alpha = 1\%$, $\beta = 0.7$) and ($\alpha = 1.5\%$, $\beta = 0.5$) have less frequent term sets, the classification results didn't decrease obviously. It indicated that the number of frequent term sets extracted by these two settings was close to the extremum but more term sets didn't affect the classification either.

Table 2. Classification results with 4 metric settings

	3:7			7:3		
Metric setting	Precision	Recall	F_1	Precision	Recall	F_1
$\alpha = 1\%$ $\beta = 0.5$	0.747	0.743	0.745	0.805	0.815	0.810
$\alpha = 1\%$ $\beta = 0.6$	0.746	0.756	0.751	0.812	0.806	0.809
$\alpha = 1\%$ $\beta = 0.7$	0.731	0.745	0.738	0.805	0.801	0.803
$\alpha = 1.5\%$ $\beta = 0.5$	0.725	0.737	0.731	0.797	0.795	0.796

Table 3 compares the classification results of extended features and the original features with the ratio of training documents to test documents as 3:7 and 7:3. And the metric setting is $\alpha = 1\%$ and $\beta = 0.6$. When the ration of training documents to test document s is 3:7, no matter single class or the overall result of precision, recall and F_1 were greatly improved by extended features. The F_1 value increased about 4.3%. When the ration of training documents to test document s is 7:3, after feature extension, the precision, recall and F_1 of News, Automobile, Female, Education and Travel were improved; the precision of Business, Sports and Technology increased but the recall decreased. On Entertainment, the precision decreased slightly but the recall increased. The reason of the different results is that when the ratio is 7:3, training documents contains more terms than 3:7 which reduced the effect of feature extension.

Table 3. Classification results of original features and extended features

	Original(3:7)			Extended(3:7)			Original(7:3)			Extended(7:3)		
Class	Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1
Automobile	0.765	0.682	0.721	0.849	0.829	0.839	0.859	0.839	0.849	0.905	0.852	0.878
Business	0.626	0.558	0.590	0.691	0.610	0.648	0.720	0.618	0.665	0.706	0.652	0.678
Technology	0.645	0.582	0.612	0.681	0.632	0.656	0.870	0.647	0.742	0.758	0.754	0.756
Education	0.863	0.912	0.887	0.896	0.918	0.907	0.870	0.922	0.895	0.923	0.935	0.929
News	0.480	0.415	0.445	0.504	0.462	0.482	0.508	0.508	0.508	0.548	0.536	0.542
Sports	0.811	0.920	0.862	0.809	0.931	0.866	0.872	0.924	0.897	0.911	0.925	0.918
Travel	0.628	0.642	0.635	0.712	0.718	0.715	0.812	0.810	0.811	0.819	0.805	0.812
Female	0.748	0.768	0.758	0.815	0.795	0.805	0.891	0.865	0.878	0.907	0.897	0.902
Entertainment	0.672	0.789	0.726	0.695	0.812	0.749	0.779	0.875	0.824	0.815	0.856	0.835
Average	0.693	0.696	0.693	0.739	0.745	0.741	0.741	0.779	0.785	0.810	0.801	0.806

To examine the classification result on different scale of training documents, we selected 7200 documents for test and respectively 3600, 7200, 10800, 14400, 18000, 21600, 25200 and 28800 documents for training, and then compare the classification result of original and extended features. Figure 1-3 are the precision, recall and F_1 on different training documents with $\alpha = 1\%$ and $\beta = 0.5$. When training documents are rare, feature extension can greatly improve precision, recall and F_1. Each of the measures increases about 3%~4%. As the number of training documents increases, the improvement margin declines. When the number of training documents is 28800, F_1 value increases about 0.8%. The result indicates that the background knowledge based on frequent term sets is an effective way for short text feature extension. When the number of training documents is rare, these extended features can greatly improve the classification results. As the number of training documents increases, although extended feature still improve the classification, more terms may compensate the shortage of features and classification accuracy didn't increase that much.

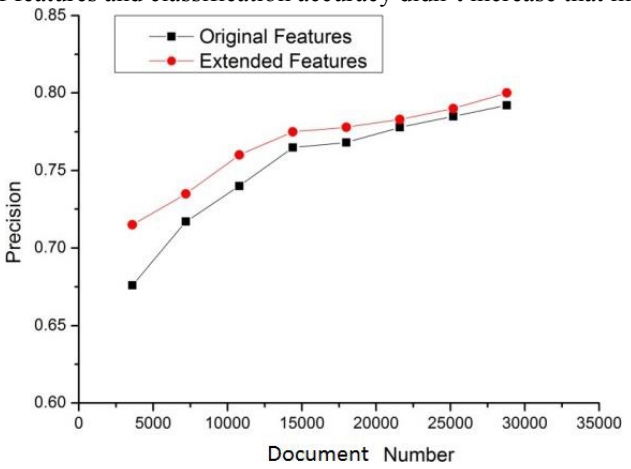

Fig. 1. Precision on different scale of training documents

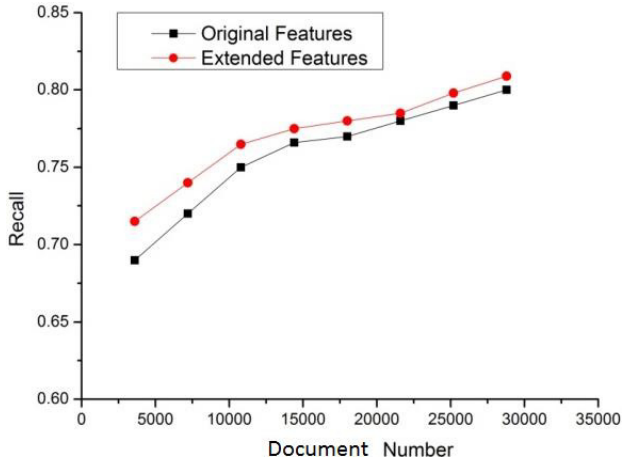

Fig. 2. Recall on different scale of training documents

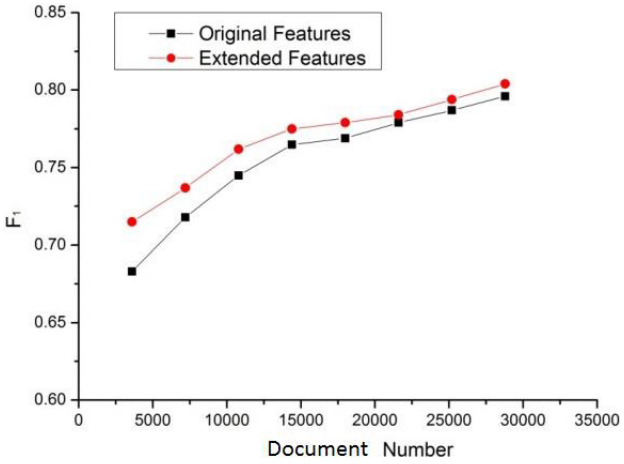

Fig. 3. F_1 on different scale of training documents

4. Conclusion

This paper discussed the short text representation problem on vector space model and proposed a feature extension method for short text categorization. Frequent term sets with identical class orientation were extracted as the background knowledge to extend the original features. For each single term of the short text, the term sets containing this term are retrieved in the background knowledge and added into the original term vector. The experimental results indicate that the support and confidence have great impact on the scale of the background knowledge, but excessive extension also has redundancy and cannot obtain further improvement. It can be concluded that the background knowledge based on frequent term sets is an effective way for short text feature extension. And when the number of training documents is rare, these extended features can greatly improve the classification results.

References

1. Gupta V, Lehal G S. A survey of text mining techniques and applications[J]. *Journal of Emerging Technologies in Web Intelligence*, 2009, 1(1): 60-76.
2. Alexander P, Patrick P. Twitter as a corpus for sentiment analysis and opinion mining[C]// *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. Valletta, Malta,2010:19-21.
3. Navigli R. Word sense disambiguation: a survey[J]. *ACM Computing Surveys*, 2009, 41(2): 1-69.
4. Zhang W, Yoshida T, Tang X. Text classification based on multi-word with support vector machine[J]. *Knowledge-Based Systems*, 2008, 21(8): 879-886.
5. Sun A. Short text classification using very few words[C]// *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. New York, USA, 2012: 1145-1146.
6. Cilibrasi R L, Vitanyi P M B. The google similarity distance[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2007, 19(3): 370-383.
7. Hu X, Zhang X, Lu C, et al. Exploiting Wikipedia as external knowledge for document clustering[C]// *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. Paris, France, 2009: 389-396.
8. Hu J, Fang L, Cao Y. Enhancing text clustering by leveraging Wikipedia semantics[C]// *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval.* Singapore, Singapore, 2008: 179-186.
9. Han J, Cheng H, Xin D, et al. Frequent pattern mining: current status and future directions[J]. *Data Mining and Knowledge Discovery*, 2007, 15(1): 55-86.

10. Cheng H, Yan X, Han J, et al. Discriminative frequent pattern analysis for effective classification[C]// *IEEE 23rd International Conference on Data Engineering*. Istanbul, Turkey, 2007: 716-725.
11. Ahonen M H. Discovery of frequent word sequences in text[C]// *the ESF Exploratory Workshop on Pattern Detection and Discovery*. London, UK, 2002: 180-189.
12. Hernández R E, García H R A, Carrasco O J A. Document Clustering Based on Maximal Frequent Sequences[J]. *Lecture Notes in Computer Science*, 2006, 4139:257-267.
13. Sogou Labs, Text Categorization Dataset [EB/OL]. (2008-09-01). http://www.sogou.com/labs/dl/c.html
14. Zhang, H.P., et al. HHMM-based Chinese lexical analyzer ICTCLAS [C]//*Proceedings of the second SIGHAN workshop on Chinese language processing*. Sapporo, Japan, 2003: 184-187.